Am. J. Hum. Genet. 49:746–756, 1991

DNA Typing and Genetic Mapping with Trimeric and Tetrameric Tandem Repeats

Al Edwards,* Andrew Civitello,* Holly A. Hammond,* and C. Thomas Caskey*,†

*Institute for Molecular Genetics and †Howard Hughes Medical Institute, Baylor College of Medicine, Houston

Summary

Tandemly reiterated sequences represent a rich source of highly polymorphic markers for genetic linkage, mapping, and personal identification. Human trimeric and tetrameric short tandem repeats (STRs) were studied for informativeness, frequency, distribution, and suitability for DNA typing and genetic mapping. The STRs were highly polymorphic and inherited stably. A STR-based multiplex PCR for personal identification is described. It features fluorescent detection of amplified products on sequencing gels, specific allele identification, simultaneous detection of independent loci, and internal size standards. Variation in allele frequencies were explored for four U.S. populations. The three STR loci (chromosomes 4, 11, and X) used in the fluorescent multiplex PCR have a combined average individualization potential of 1/500 individuals. STR loci appear common, being found every 300–500 kb on the X chromosome. The combined frequency of polymorphic trimeric and tetrameric STRs could be as high as 1 locus/20 kb. The markers should be useful for genetic mapping, as they are sequence based, and can be multiplexed with the PCR. A method enabling rapid localization of STRs and determination of their flanking DNA sequences was developed, thus simplifying the identification of polymorphic STR loci. The ease by which STRs may be identified, as well as their genetic and physical mapping utility, give them the properties of useful sequence tagged sites (STSs) for the human genome initiative.

Introduction

The study of satellite DNA (Brutlag 1980), VNTRs (Nakamura et al. 1987), and evolutionary comparisons of orthologous DNA sequence (Savatier et al. 1985) had suggested that tandemly repeated DNA is frequently polymorphic. Variation at dimeric [AC] short tandem repeats (STRs) or microsatellites is now well known (Litt and Luty 1989; Weber and May 1989). Tandemly repeated degenerations of the poly(A) tail of *Alu* repeats have also been found to be polymorphic (Dryja et al. 1989; Economou et al. 1990; Orita et al. 1990; Sinnett et al. 1990). We have found trimeric and tetrameric STRs to be highly polymorphic in humans (Edwards et al. 1989) and to be easily amplified with the PCR (Saiki et al. 1988). Allele lengths of STRs are precisely resolved to single bases by analysis of 100–500-bp amplified fragments on polyacrylamide sequencing gels. The high variability of STRs and the ease with which they may be used as polymorphic markers renders them suitable for application to the human genome initiative and other projects requiring a large number of physical and genetic markers. The precision of sequence-based alleles has the potential to improve on the limitations of continuous-allele systems (e.g., VNTRs) currently in use for genetic typing in forensic science and medicine (Devlin et al. 1990).

We report the results of studies designed to (1) allow us to understand the distribution and frequency of trimeric and tetrameric STRs in the human genome, (2) extend our original observation that trimeric and tetrameric STRs are highly variable and may amplify more faithfully than $[AC]_n$ repeats to additional loci, (3) demonstrate the feasibility of a DNA typing (fingerprinting) assay with internal size standards based

Received March 26, 1991; revision received June 10, 1991.
Address for correspondence and reprints: Al Edwards, Ph.D., Institute for Molecular Genetics, Baylor College of Medicine, One Baylor Plaza, T809, Houston, TX 77030.

0002-9297/91/4904-0007$02.00

on multiplex PCR of STRs coupled to a fluorescent detection system, and (4) develop a general strategy for obtaining the DNA sequence flanking a STR locus in any cloned DNA segment. The data indicate that trimeric and tetrameric STRs are frequent in the human genome, amplify faithfully, and are suitable to application to DNA typing and genetic mapping.

Subjects and Methods

Population Groups

DNA was prepared (Patel et al. 1984) from unrelated volunteer donors at a local blood bank. Blood donors were visually categorized as white, black, and other by blood bank personnel. Hispanic and Asian samples were identified within the "other" group on the basis of common surnames in the Houston area. Thus, the population samples are representative of four large ethnic groups in the Houston metropolitan area. For each population group the number of individuals studied is given in table 3.

Genetic Loci

The genetic loci studied are listed in table 1. The loci are named with reference to their GenBank locus designations and the lowest alphabetical representation of the STR motif. For example, HUMHPRTB[AGAT]$_n$, is the polymorphic [AGAT] repeat located within the DNA sequence of the GenBank entry HUMHPRTB (the human HPRT genomic sequence). The human gene mapping (HGM) symbols and GenBank accession numbers are as follows: FABP2, M18079 and J03465; PLA2, M22970 and M14965; AR, M21748; EPO, X02158; TH, D00269; TNFB, M16441; REN, M10151; and HPRT, M26434. HUMSTRX1 was submitted under accession number M38419, and HUMGPP33A09 has accession numbers M32647 and J05427. The number of tandem repeats was determined by direct DNA sequencing or by subcloning and sequencing the amplified products from at least two alleles (author's unpublished data). DNA loci were identified either (1) by a pattern search of GenBank DNA sequences with STR motifs by using the pattern2 program available on the Molecular Biology Information Resource (MBIR) here at Baylor (Dr. Charlie Lawrence, Department of Cell Biology) or (2) with the STR-PCR method described herein.

Genotype Determinations

Genotypes of individuals from the four population groups were determined with multiplex PCRs coupled to either a radioactive or fluorescent detection system. Oligonucleotides for the PCRs were synthesized on an Applied Biosystems (ABI) 380B DNA synthesizer. Underivatized oligonucleotides were not purified after deprotection and lyophilization. ABI Aminolink 2 chemistry was used to derivatize oligonucleotides for biotin and fluorescent labeling, after which they were ethanol precipitated and purified by PAGE.

Radioactive detection of single and multiplex PCR products was achieved by including 2–4 μCi ^{32}P-α-dCTP and the desired primer sets in a standard reaction cocktail (Saiki et al. 1988). Amplification conditions were as described below for the fluorescent multiplex DNA typing assay. The radioactively labeled products were separated by electrophoresis through a standard denaturing DNA-sequencing gel, after which the gel was fixed, dried, and used to expose film. Alleles were identified relative to control DNA samples of known size. Controls without DNA were run with each set of experiments (cocktail). The primers were as follows (5′→3′): HUMFABP[AAT]$_n$ gtagtatcagtttcatagggtcacc and cagttcgtttccattgtctgtccg; HUMARA[AGC]$_n$, tccagaatctgttccagagcgtgc and gctgtgaaggttgctgttcctcat; HUMGPP3A09[AAGG]$_n$, tgtgagtcccagttgccagtctac and actggtcaccttggaaagtggcat; HUMERP[c:AATG,ACTC,c:AATG]$_n$, tgagggctgtatggaatacgttca and caagcaccaagctgagcaaacaga; HUMTH01[AATG]$_n$, gtgggctgaaaagctcccgattat and attcaaagggtatctgggctctgg; HUMTNFAB[AATG]$_n$, ggagagacaggatgtctggcacat and ccatctctccttagctgtcata; HUMRENA4[ACAG]$_n$, agagtaccttccctcctctactca and ctctatggagctggtagaacctga; HUMHPRTB[AGAT]$_n$, atgccacagataatacacatcccc and ctctccagaatagttagatgtagg; HUMSTRX1[AGAT]$_n$, ctccttgtggccttccttaaatgg and cttctccagcacccaaggaagtca; and HUMPLA2A1[AAT]$_n$, ggttgtaagctccatgaggttaga and gtcctaggagctagagatacagc. These methods have been described in detail elsewhere (Edwards et al., submitted).

The fluorescent dyes (Molecular Probes, Eugene, OR) used in the fluorescent multiplex assay were (1) NBD hexanoic acid (green pen) for all internal standard markers, (2) 5-(and-6-)-carboxyfluorescein succinimidyl ester (blue pen) for the HUMTH01[AATG]$_n$ and HUMHPRTB[AGAT]$_n$ loci, and (3) Texas Red™ sulfonyl chloride (red pen) for the HUMFABP[AAT]$_n$ locus. The primer sets (5′→3′) were as follows, with the first being derivatized and the second containing an *Mlu*I restriction site: HUMTH01[AATG]$_n$ gtgggctgaaaagctcccgattat and ttacgcgtattcaaagggtatctgggctctgg; HUMFABP[AAT]$_n$, gtagtatcagtttcatagggtcacc and ttacgcgtctcggacagtattcagttcgtttc; and HUM-

HPRTB[AGAT]$_n$, atgccacagataatacacatcccc and ttacgcgttctccagaatagttagatgtaggtat. Simultaneous amplification with all six primers was performed with 25 cycles by denaturing at 95°C for 45 s, annealing at 60°C for 30 s, and extending at 72°C for 30 s, by using Perkin Elmer—Cetus thermocyclers, AmpliTaq, and buffer conditions. The concentration of primers in the multiplex were 0.06 μM for HUMTH01[AATG]$_n$, 1.6 μM for HUMFABP[AAT]$_n$, and 0.56 μM for HUMHPRTB[AGAT]$_n$. After amplification, the products were phenol extracted, ethanol precipitated, and digested with *Mlu*I. The digested multiplex products were then combined with internal size standards and electrophoresed through a 6.5% polyacrylamide, 8.3 M urea gel at 1,300–1,500 V, 24 mA, and 32 W, at a temperature of 46°C. Internal size standards were prepared by amplifying specific alleles from individuals of known genotype. The products were quantitated, combined to give near equimolarity, diluted approximately 5,000-fold, and reamplified with approximately 12 cycles.

Plaque Hybridizations

The frequency of specific STR sequences in the human genome and X chromosome was estimated by hybridization of STR core oligonucleotides to recombinant bacteriophage lambda clones transferred to a grid (Sambrook et al. 1989) from either (1) a genomic library prepared from total human genomic DNA or (2) an X chromosome–specific (sorted) bacteriophage λ library (LAOXNL01 57750) from the American Type Culture Collection (Rockville, MD). Oligonucleotides (30 bp) were end-labeled and hybridized in 5 × SSC as described elsewhere (Davis et al. 1986). The oligonucleotides used are shown in table 2. The requirements for a repeat target approximately the length of the core oligonucleotide was demonstrated for [AGAT] by washing at increasing stringencies. Further, specificity was achieved in that STR core oligonucleotides differing by one base in the repeat motif (e.g., [AATC] and [AATG]) hybridized to different clones. A total of 1,020 recombinant bacteriophage were hybridized to radiolabeled 30-bp oligonucleotides, and calculations were based on an average insert size of 15 kb in the library.

STR-PCR

The oligonucleotide linker for anchored PCR described by Riley et al. (1990) was modified to enable amplification of the DNA sequence flanking a STR in any cloned DNA segment (fig. 3). The oligonucleotides used were actgcagagacgctgtctgtcgaaggtaaggaacggacgagagaagggagag and ctctcccttctcgaatcgtaaccgttcgtacgagaatcgctgtctctgcagt, which, when annealed to each other, generate a linker with the following features: (1) blunt ends to enable ligation to any blunt-ended DNA molecule, (2) a priming site for a DNA sequencing primer (tacgagaatcgctgtctctgcagt), and (3) a region of noncomplementarity (bubble) in which one strand is the same sequence as the anchor PCR primer (gccggatcccgaatcgtaaccgttcgtacgagaatcgc). The oligonucleotide not containing the anchor PCR primer sequence was phosphorylated (Sambrook et al. 1989), and the linker formed by annealing the two oligonucleotides. Blunt ends were created in cloned DNA containing STRs, by digestion with *Alu*I, *Hae*III, or *Rsa*I. The linker was ligated to the cloned DNA in a standard DNA ligation cocktail containing 8.3 ng digest DNA/μl and 0.28 μg linker/μl. Amplification was performed with the specific internal STR-PCR primer ([AGAT]$_{7.5}$ or [ATCT]$_{7.5}$) and the anchor PCR primer as described above, except that the concentration of Mg^{++} was 1 mM, the extension time was 1 min at 72°C and the number of cycles was 25–30.

The biotinylated amplification products (25 μl) were captured with an equal volume of avidin-coated magnetic beads (M-280), obtained from Dynal, by incubation on a wheel for 30 min at 25°C. The supernatant was removed, and the nonbiotinylated strand was eluted by denaturation with 150 μl 0.15 M NaOH for exactly 5 min. The beads were washed twice with water and resuspended in 7 μl water for DNA sequencing using the Sequenase Kit (United States Biochemical).

Calculations

Unbiased estimates of the heterozygote frequency were obtained from the genotype data of unrelated individuals, as described elsewhere (Chakraborty 1990). Individualization potentials were calculated using a standard formula (Sensabaugh 1982)—namely, the sum of the squares of all possible genotype frequencies.

Results and Discussion

STR Loci Studied and Genomic Frequency

Eighteen STR loci were studied for polymorphism. Opposing oligonucleotide primers flanking the STRs were used to amplify across each locus in the presence of approximately 2 μC ^{32}P-α-dCTP, and the products

Table 1

Polymorphic STR Loci

Locus and STR Sequence	No. of Alleles Detected	No. of Chromosomes Studied	Chromosomal Location[a]
HUMFABP$[AAT]_{8-15}$	8	314	4
HUMPLA2A$[AAT]_{16}$	7	300	12
HUMARA$[AGC]_{12-30}$	17	228	X
HUMGPP3A09$[AAGG]_{9}$	4	24	ND
HUMERP$[c{:}AATG]_{2}[ACTC]_{4}[c{:}AATG]_{5}$	2	20	7
HUMTH01$[AATG]_{6-12}$	7	320	11
HUMTNFAB$[AATG]_{5}$	4	24	6
HUMRENA4$[ACAG]_{7-12}$	6	310	1
HUMHPRTB$[AGAT]_{9-16}$	8	227	X
HUMSTRX1$[AGAT]_{13}$	11	44	X

[a] ND = not determined.

were analyzed on DNA sequencing gels. Ten STR loci were polymorphic (table 1). Mendelian inheritance of alleles was established by study of more than 25 meioses (author's unpublished data). The intensity of artifactual background bands generated during amplification varied from undetectable to approximately 5% of the adjacent allele, enabling precise length determination in each case. The longer and more complex trimeric and tetrameric STR motif may account for the greater amplification fidelity than we have observed for many [AC] repeats reported in the literature.

We observed only one allele for the following eight loci in the 20 unrelated chromosomes studied: HUMVTNR$[AAC]_{10}$, HUMCENPBR$[ACC]_{27}$, HUMKER671$[ACC]_{19}$, HUMAK1$[AGC]_{7}$, HUMPLAP1B$[AGC]_{7}$, HUMHPRTB$(A)_{2-5}C]_{11}$, HUMHPRTB$[AATC]_{7}$, and HUMFIXG$[AATT]_{5}$. Comparison of the polymorphic loci with the monomorphic loci suggests that the actual sequence of the core repeat motif is not the primary factor in the determination of polymorphism. These studies suggest that tandem reiteration, regardless of the core sequence motif, predisposes to variation but is not the exclusive factor. Increasing numbers of tandem repeats at a locus have been reported to be associated with the informativeness of dimeric [AC] STRs (Weber 1990). The same trend is observed with the trimeric and tetrameric STRs for which we have heterozygote frequencies (table 1).

STR loci found in the primate sequences of GenBank indicate that trimeric and tetrameric STRs are found within coding, genic (introns and flanking sequences), and extragenic regions of the genome (author's unpublished data). As might be expected from the genetic code, only trimeric STRs were found within coding sequences. The polymorphic $[AGC]_n$ STR is found within the protein coding region of the human androgen receptor gene (table 1) and within the *Drosophila Notch* gene (Wharton et al. 1985; Tautz 1989). In both cases the translational frame of the STR encodes glutamine repeats. Glutamine repeats are also found in other receptors involved in growth and development (Wharton et al. 1985). The role of such repeats in receptor function is unknown. Variable tandem reiteration of large coding segments has been reported for the apolipoprotein (A) (Koschinsky et al. 1990; Lindahl et al. 1990) and PUM mucin-type glycoprotein loci (Swallow et al. 1987). Thus extensive variation of tandem repeats in coding sequences can be tolerated in some proteins.

We have explored the frequency of STRs in the total human genome and on the X chromosome. Radiolabeled STR oligonucleotides hybridized to 3%–5% of recombinant bacteriophage lambda clones in a human X chromosome library (table 2). The STRs occur on the X chromosome at a frequency of approximately 1 STR/300–500 kb. The length of the STR required to obtain a hybridization signal was approximately 7 units for the tetrameric repeats, which is of sufficient length to anticipate that at least 50% are polymorphic (table 1). Similar frequencies were obtained for the frequency of STRs in the total human genomic library. These data and dispersion of STR sequences in GenBank suggest that the trimeric and tetrameric STRs occur throughout the genome—on many, if not all, chromosomes. The A-rich STRs—e.g., [AAAG] and

Table 2

Frequency of Trimeric and Tetrameric STRs on Human X Chromosome and in Total Human Genome

STR	% Positive Bacteriophage: X Chromosome	% Positive Bacteriophage: Genome[a]	Frequency (kb/STR)
[AAT]	5	7	200–300
[AATC]	5	9	150–300
[AATG]	3	ND	500
[ACAG]	3	3	500
[AGAT]	3	4	400–500

[a] ND = not determined.

[AAAT]— often found as polymorphic tandem degenerations of the poly(A) tracts of *Alu* repeats (Dryja et al. 1989; Economou et al. 1990; Orita et al. 1990; Sinnett et al. 1990; Zuliani and Hobbs 1990), which appears to increase their frequency. Some apparently chromosome-specific, satellite-like configurations of longer STRs have been reported elsewhere (Nanda et al. 1990).

Population Genetics

The genotypes of 40 individuals in four U.S. population groups were determined in order to study the interpopulation variation of STRs (table 3). Unbiased

Table 3

Heterozygote Frequencies and Individualization in Four Population Groups

STR Locus and Population	No. of Alleles Observed	No. of Chromosome Studied	P_I[a]	H[b]
HUMFABP$[AAT]_{8-15}$:				
Black	8	80	.09	.78
White	6	76	.16	.69
Hispanic	7	80	.19	.63
Asian	5	78	.28	.52
HUMPLA2A$[AAT]_{16}$:				
Black	7	70	.05	.84
White	6	76	.13	.70
Hispanic	7	80	.08	.79
Asian	6	74	.08	.79
HUMARA$[AGC]_{12-30}$:				
Black	14	62	.02	.89
White	10	59	.01	.87
Hispanic	13	54	.01	.91
Asian	12	53	.01	.89
HUMTH01$[AATG]_{6-12}$:				
Black	5	80	.09	.78
White	6	80	.10	.76
Hispanic	5	80	.10	.77
Asian	7	80	.13	.71
HUMRENA4$[ACAG]_{7-12}$:				
Black	6	80	.31	.47
White	4	76	.47	.34
Hispanic	4	80	.39	.40
Asian	4	74	.32	.48
HUMHPRTB$[AGAT]_{9-16}$:				
Black	8	62	.09	.78
White	7	59	.13	.73
Hispanic	5	55	.16	.69
Asian	5	51	.11	.75

[a] Individualization potential.
[b] Unbiased estimate of heterozygote frequency.

heterozygote frequencies varied considerably between loci, ranging from .34 for HUMRENA4[ACAG]$_n$ to .91 for HUMARA[AGC]$_n$. Although the heterozygote frequencies between population groups tend to be similar, there are exceptions—e.g., the .26 difference, between blacks and Asians, in heterozygosity for the HUMFABP[AAT]$_n$ locus. Linkage disequilibrium between the STR loci was not detected, and the alleles from each locus appear to be in Hardy-Weinberg equilibrium (Edwards et al., submitted). Thus, the population data base may be used to calculate accurate genotype frequencies for personal identification.

The observed allele frequencies were used to calculate the probability—called the individualization potential, or P_I—that two individuals selected at random from a population would be identical for a marker (Sensabaugh 1982). Together, the five loci shown in table 3 would be expected to distinguish, on average, 1/90,000 persons. The ease with which additional loci may be identified and studied (see below) indicates that P_I's in excess of $1/10^6$ will be easily obtainable.

The frequencies of alleles from the STR loci revealed both unimodal and bimodal allele distributions, along with differences between the frequencies of specific alleles in different populations (fig. 1). The mechanism by which the different distribution modes arose has not been ascertained. The different modalities may represent the evolutionary primacy of the major alleles with mutation-generating adjacent alleles. This model is consistent with a study in which the frequency for a 2-bp deletion was 18-fold greater than that for a 4-bp deletion for an [AC] STR cloned into *Escherichia coli* (Levinson and Gutman 1987). It may be possible to identify the mechanisms of polymorphism development and generation of the allele frequency distributions by studying the evolutionary lineages of the alleles. We have amplified the HUMHPRTB[AGAT]$_n$ and HUMFABP[AAT]$_n$ loci from the baboon, gibbon, orangutan, chimpanzee, and African green monkey, thereby indicating the feasibility of such studies (author's unpublished data).

Fluorescent, Multiplex PCR Genetic Typing Assay

DNA typing is a powerful technique for determining the relationship between two genomic DNA samples. Applications for DNA typing include (1) personal identification in paternity testing and forensic science and (2) sample-source determinations in transplantation, prenatal diagnosis, and pedigree validation. Several features of polymorphic trimeric and tetrameric STRs suggest that they could form the basis of a powerful and simple DNA typing assay. The small size of the amplified units suggests that several loci could easily be amplified simultaneously by multiplex PCR (Chamberlain et al. 1988)—and be analyzed with precise allele identification on DNA sequencing gels. The precision, sensitivity, and speed of detecting alleles with the PCR offers investigative opportunities in the study of forensic specimens. The fidelity of amplification of the trimeric and tetrameric STRs indicates that the genotyping fingerprints would be easily interpreted and amenable to automation. The markers described herein have been applied to degraded DNA and adjudicated-case DNA samples (data not shown). Recent advances in fluorescent DNA fragment detection can be used for internal size standards and precise allele quantitation.

For genetic typing with internal standards, alleles from three chromosomally unlinked STR loci were amplified simultaneously in a multiplex PCR (fig. 2). One primer from each of the three amplification primer sets was differentially labeled with one of the four fluorescent dyes used with the DNA sequencing device. One dye was reserved for the internal standards, while three dyes were available for the amplification products of STR loci. In principle, any given region of the sequencing gel could contain internal standards, as well as alleles from three unlinked STR loci. Used to full potential, the approach has enormous personal identification power of high accuracy.

In each product in the multiplex PCR, amplification incorporated a fluorescent label into one end and a *MluI* site into the other end (fig. 2). Following amplification of the STR loci from a genomic DNA sample, residual activity of the *Thermus aquaticus* polymerase was destroyed, and a homogeneous fragment length was achieved for each allele by digestion with *MluI*. This step eliminates the observation of two bands separated by 1 bp, for strands (alleles) in which the addition of a single base pair by the *T. aquaticus* DNA polymerase to the 3′ end is incomplete (Clark 1988). The treated multiplex products were then mixed with internal standards and loaded onto a sequencing gel for analysis on an ABI 370A (Smith et al. 1986).

Internal standards were generated by pooling amplification products from individuals of known genotype, such that the molar ratios of each allele observed were approximately equal. The pooled alleles were diluted, reamplified, and treated with *MluI*. This scheme for generating internal standard size markers insures a virtually unlimited supply of standards.

The combination of a quantitative detection system

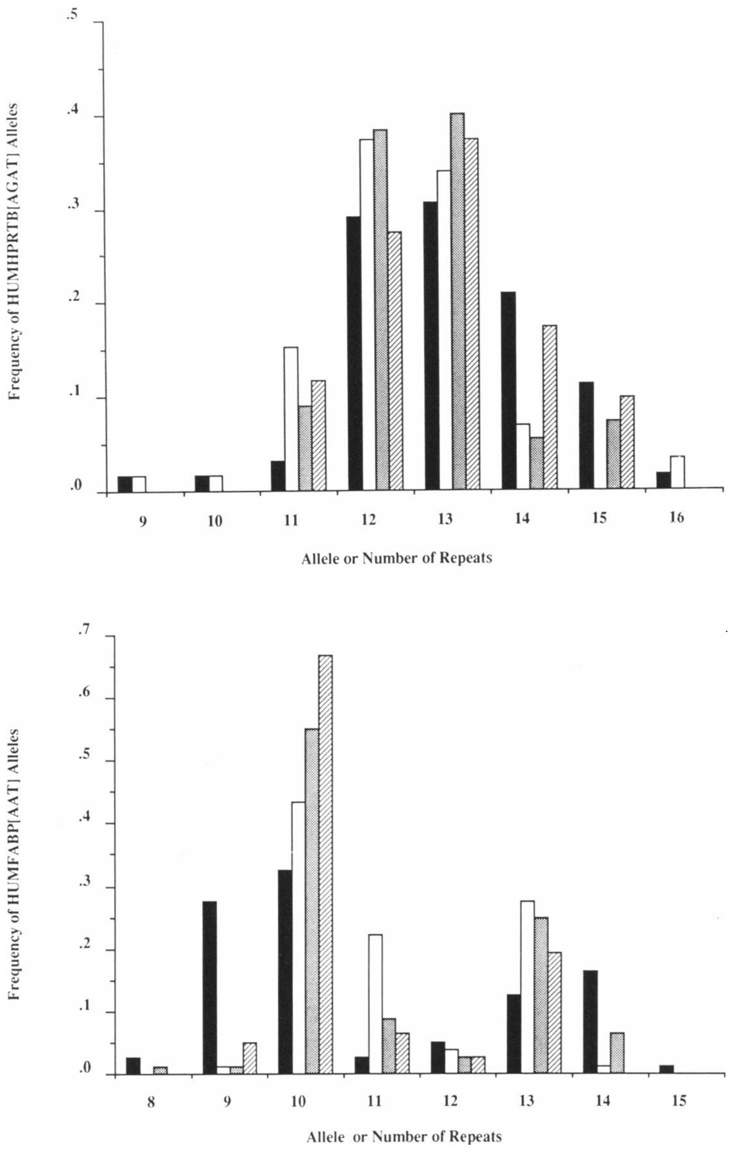

Figure 1 Frequencies of alleles at HUMHPRTB[AGAT]$_{9-16}$ (*top*) and HUMFABP[AAT]$_{8-15}$ (*bottom*) loci in four population groups. The genotypes of approximately 40 individuals in each group were determined. The alleles are numbered as in table 1. The total number of chromosomes examined is given in table 2. ■ = black; □ = white; ▩ = Hispanic; ▨ = Asian.

and multiplex PCR enables additional levels of internal control and precision. By use of multiplex PCR products synthesized under standardized amplification conditions, it may be possible to relate the fluorescent intensity of specific alleles at different loci. As shown in figure 2, the relationship between alleles of different loci makes it possible to distinguish between homozygosity and hemizygosity at a given locus. Determining the extent to which this observation will be valid in practice will require further study. While failure of allele amplification could occur by primer binding-site mutation (not yet observed), the null would be detectable by quantitation. This quantitative capacity and the high resolution of sequencing gels removes the doubt which has been cast on the use of VNTRs by the observation of homozygosity excess in population studies (Devlin et al. 1990). The quantitative nature of the allele identification also may facili-

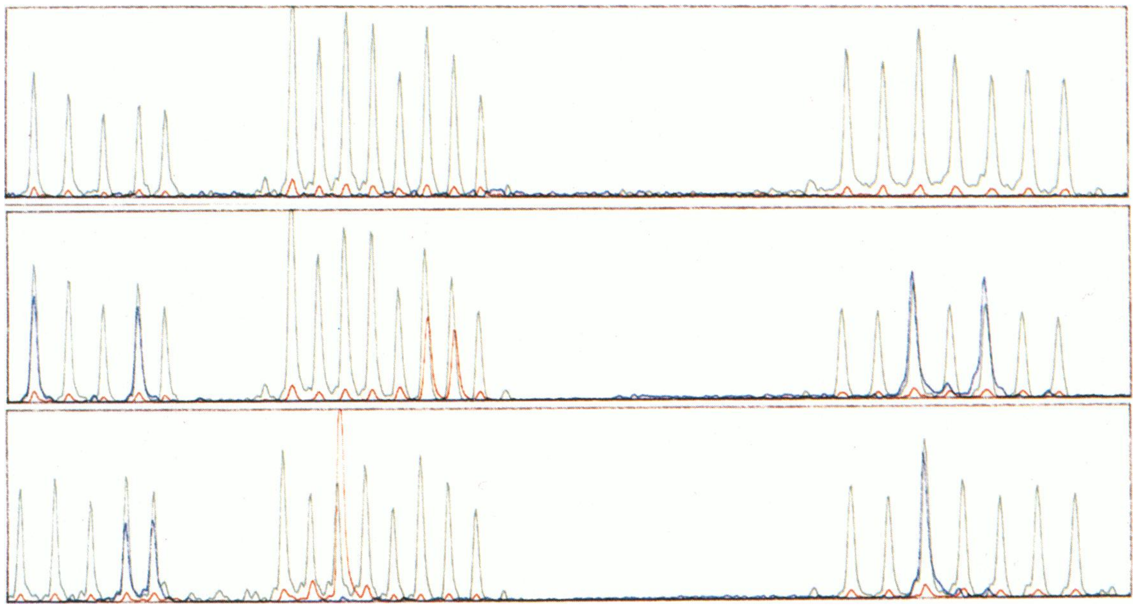

Figure 2 Fluorescent DNA typing assay. *Top,* Fluorescent profiles of internal standard cocktails when combined and electrophoresed in single lane of ABI 370A DNA sequencing device. *Middle,* Internal standards combined with amplification products of multiplex PCR composed of (left to right) HUMTH01[AATG]$_n$ (blue), HUMFABP[AAT]$_n$ (red), and HUMHPRTB[AGAT]$_n$ (blue) loci. The individual shown is heterozygous for all three markers. *Bottom,* Multiplex amplification from individual homozygous at HUMFABP[AAT]$_n$ and hemizygous at HUMHPRTB[AGAT]$_n$. The analysis software (Mapper) scales the intensities of the fluorescent profiles relative to the strongest signal.

tate the analysis of mixed body samples in forensics and prenatal diagnosis. The method may have application to the detection of chromosomal aneuploidy and somatic mosaicism seen in patients with chromosomal abnormalities and following bone marrow transplantation.

The combined P_I of the three loci together is 1/500 individuals. The combined genotype frequencies (three loci) of the individuals in panels A and B are 1/3,850 and 1/120, under Hardy-Weinberg equilibrium. The addition of three more loci would give a P_I of approximately 1/200,000, while the addition of six more loci would give a P_I of $1/9 \times 10^7$. Multiplex PCRs of this complexity are feasible, as we have shown with the eight- and nine-genetic-site multiplexes for the hypoxanthine phosphoribosyltransferase (Gibbs et al. 1990) and dystrophin genes (Chamberlain et al. 1988).

Rapid Study of STR Loci

The frequency of particular STRs is sufficiently high that it should be possible to identify a STR locus in most genomic segments, say 20 kb, via screening by hybridization with a limited set of STR oligonucleotides. Although screening by hybridization may be performed easily, obtaining the DNA sequence flanking both sides of the STR is laborious. We have designed an anchored-PCR-based method to amplify the flanking segments and to determine their DNA sequence by solid-phase DNA sequencing.

We adapted a published anchored-PCR method (Riley et al. 1990) for amplification of the DNA sequences flanking a STR locus (fig. 3). Only by synthesis from the internal specific primer (STR primer) is the priming site for the anchored PCR primer created (fig. 3). The amplified products were then sequenced after capture of the appropriate strand via a biotin/avidin solid-support technology. The STR-PCR method has been applied with success to five recombinant bacteriophage lambda phage clones which hybridized to the radiolabeled [AGAT] oligonucleotide. Figure 4 shows results from two bacteriophage clones. In the case of the clone derived from the HPRT locus, the products were of the molecular weight expected on

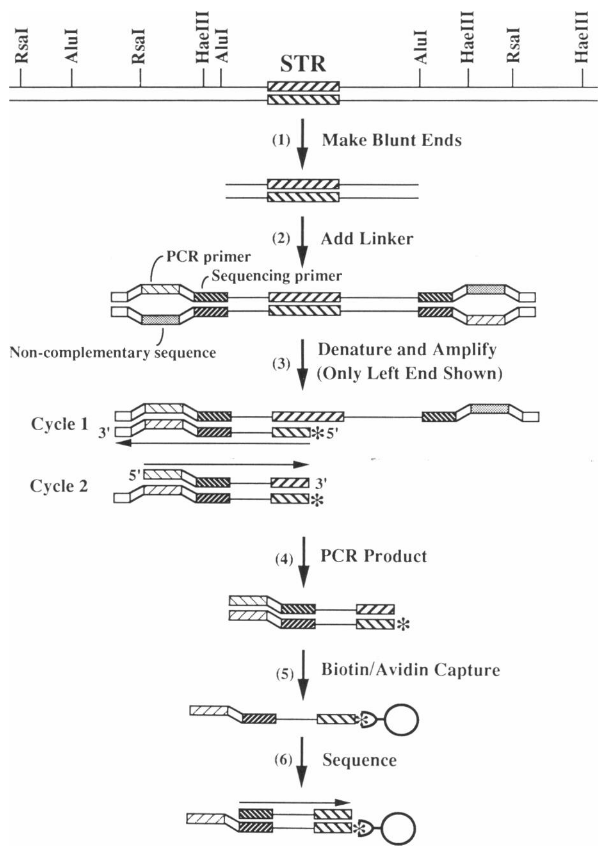

Figure 3 Strategy to determine sequence flanking STR. (1), Blunt ends generated flanking both sides of STR in cloned DNA segment by digestion with single restriction enzyme. Multiple enzymes may be used, separately, to generate a flanking sequence length in the range of 100–1,500 bp. (2), Linker similar to that of Riley et al. (1990), which contains region of noncomplementary DNA and which is ligated to population of blunt-ended molecules. (3). Flanking sequences amplified in separate reactions. The left end is amplified with the anchored PCR primer and a primer of one strand of the STR. The right end is amplified with the same anchored PCR primer and a primer of the other strand of the STR. The STR primers may be biotinylated (*) for the following steps. (4). Final biotinylated (*) PCR product. (5), Biotinylated strand, which may be captured with avidin-coated beads. (6), Flanking sequence, which may be obtained by extension from sequencing primer in presence of dideoxynucleotides.

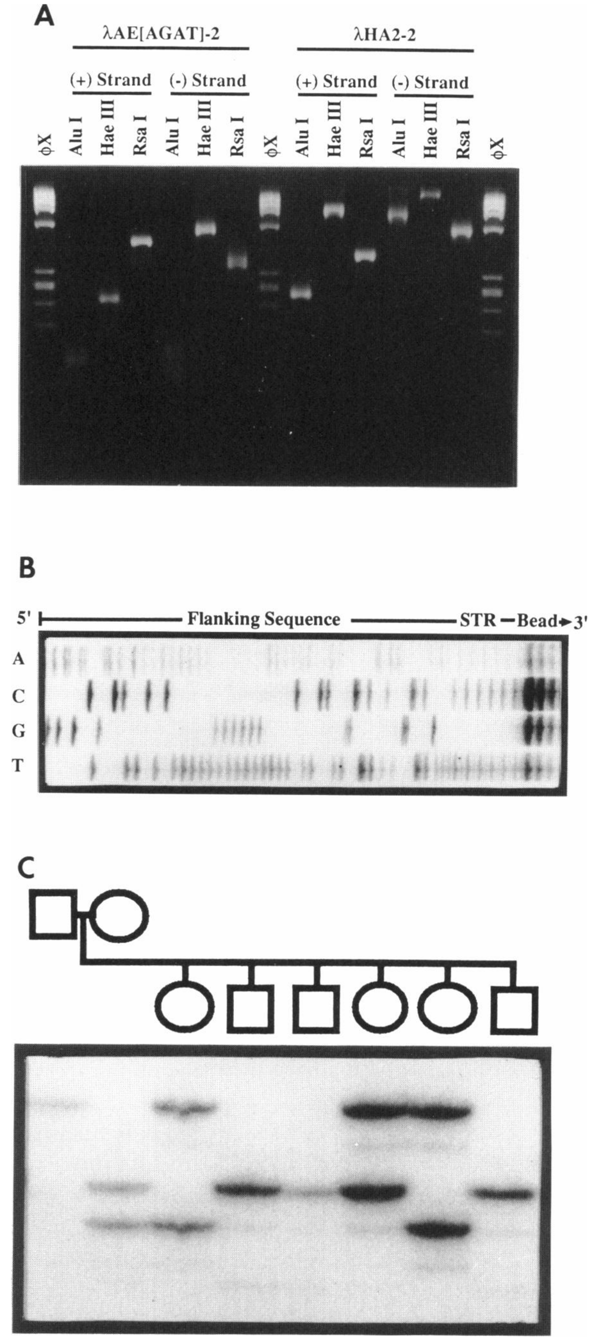

Figure 4 Development of polymorphic STR locus. See fig. 3 for strategy. *Top,* Amplification of DNA sequence flanking both sides of [AGAT] STR from two recombinant bacteriophage. *Middle,* Direct DNA sequencing of single-stranded template, following capture and strand separation of biotinylated amplification products of λAE[AGAT]-2 with avidin-coated magnetic beads. *Bottom,* Oligonucleotides complementary to sequence flanking STR that were used to amplify STRX1 locus in family.

the basis of the published sequence (Edwards et al. 1990). An anonymous clone, HUMSTRX1[AGAT]$_n$, was isolated from the X chromosome library and was found to contain 13 tandem reiterations of [AGAT]. HUMSTRX1[AGAT]$_n$ was highly variable (table 1).

The ability to determine easily the sequence of DNA segments flanking STR loci in cloned fragments should aid investigators seeking polymorphic markers in their region of interest. It may be possible to extend the method to more complex sources, such as yeast artificial chromosomes or genomic DNA. In these cases, it might be necessary to subclone and sequence the amplification products, depending on the number of amplified products. The flanking sequence opposing a sequenced flanking segment could be obtained by amplification across the STR; the PCR would amplify between a primer derived from the sequenced end (e.g., the left side) and the anchor on the opposing end (i.e., the right end). DNA sequencing of this amplified product would provide the second primer for the development of a PCR assay for the marker.

Conclusions

We have studied 18 trimeric and tetrameric STRs and have found approximately half to be polymorphic. Loci with greater numbers of repeats are more likely to be polymorphic (Weber 1990). The STR loci are associated with high heterozygosities in the human population, with notable differences between population groups in heterozygosities and allele frequencies. A DNA typing assay based on multiplex PCR amplification of STR loci is described which features internal standards for precise allele identification and fluorescent detection for quantitation. Application of DNA typing to paternity testing, forensic identification, medical diagnosis, and genetic studies will be facilitated by the ability to determine precise allele sizes and to distinguish between real and apparent homozygotes. The amplification of the trimeric and tetrameric STRs appears to be more faithful than that which has been found with dimeric $[AC]_n$ repeats.

The combined frequency for all 44 possible unique trimeric and tetrameric STRs in the genome is estimated at 400,000 or 1 STR (of approximately seven repeat units in length)/10 kb. A method is described by which DNA segments flanking a STR locus may be amplified and sequenced directly. After oligonucleotides flanking the STR are synthesized, the locus may be studied for variation by using the PCR.

The clarity with which alleles of the highly polymorphic and frequent trimeric and tetrameric STRs can be scored makes them ideal genetic markers. Our studies of the CEPH families with $HUMHPRTB[AGAT]_n$ and $HUMFABP[AAT]_n$ indicate that such STRs are reliable polymorphisms (Edwards et al., submitted). The power of STRs as genetic mapping tools has recently been illustrated in two circumstances. After exclusion of facioscapulohumoral dystrophy from 90% of the genome by conventional probe RFLPs had indicated the need for additional markers, linkage to a PCR-based dimeric STR on chromosome 4 was achieved (Wijmenga et al. 1990). Linkage, in a small family, of an X-linked mental retardation locus to the $HUMHPRTB[AGAT]_n$ STR demonstrates the potential for disease-locus identification in small kindreds which have proved refractory to study with RFLPs (Huang et al., submitted). It should be possible to construct sets of multiplex PCRs composed of STRs from specific chromosomes, for rapid linkage first to chromosomal bands and then sequentially to diminishing physical distances. These markers may also be used for the physical mapping of yeast artificial chromosomes and for correlation of the genetic and physical maps.

Acknowledgments

We thank Drs. Ranajit Chakraborty and Belinda J. F. Rossiter for criticisms; Drs. Raymond Fenwick, Xiangwei Wu, Derek Kuhl, and Mark Tristan for reagents; and Dr. Mal Raff of Applied Biosystems, Inc. for the Mapper Software. A.E. is a Josephine and Edward Hudson Scholar and Medical Scientist Training Fellow; C.T.C. is an investigator with the Howard Hughes Medical Institute. This paper was prepared under grant 90-IJ-CX0037 from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. Points of view or opinions in this document are those of the author(s) and do not necessarily represent the official position or policies of the U.S. Department of Justice.

References

Brutlag DL (1980) Molecular arrangement and evolution of heterochromatic DNA. Annu Rev Genet 14:121–144

Chakraborty R (1990) Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. Am J Hum Genet 47:87–94

Chamberlain JS, Gibbs RA, Rainer JE, Nguyen P-N, Caskey CT (1988) Deletion screening of the Duchenne muscular dystrophy locus *via* multiplex DNA amplification. Nucleic Acids Res 16:11141–11156

Clark JM (1988) Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic DNA polymerases. Nucleic Acids Res 16:9677–9686

Davis LG, Dibner MD, Battey JF (1986) Hybridization with synthetic ^{32}P end-labeled probe. In: Davis LG, Dibner MD, Battey JF (eds) Methods in molecular biology. Elsevier Science, New York, pp 75–78

Devlin B, Risch N, Roeder K (1990) No excess of homozygosity at loci used for DNA fingerprinting. Science 249: 1416–1420

Dryja TP, Mukai S, Peterson R, Rapaport JM, Walton D, Yandell DW (1989) Parental origin of mutations of the retinoblastoma gene. Nature 339:556–558

Economou EP, Bergen AW, Warren AC, Antonarakis SE (1990) The polydeoxyadenylate tract of *Alu* repetitive elements is polymorphic in the human genome. Proc Natl Acad Sci USA 87:2951–2954

Edwards A, Gibbs RA, Nguyen PN, Ansorge W, Caskey CT (1989) Automated DNA sequencing methods for detection and analysis of mutations: applications to the Lesch-Nyhan syndrome. Trans Assoc Am Physicians 102: 185–194

Edwards A, Hammond HA, Caskey CT, Chakraborty R. Population genetics of trimeric and tetrameric tandem repeats in four human ethnic groups (submitted)

Edwards A, Voss H, Rice P, Civitello A, Stegemann J, Schwager C, Zimmermann J, et al (1990) Automated DNA sequencing of the human HPRT locus. Genomics 6:593–608

Gibbs RA, Nguyen PN, Edwards A, Civitello AB, Caskey CT (1990) Multiplex DNA deletion detection and exon sequencing of the hypoxanthine phosphoribosyltransferase gene in Lesch-Nyhan families. Genomics 7:235–244

Huang TH-M, Hejtmancik JF, Edwards A, Pettigrew L, Herrera CA, Hammond HA, Caskey CT, et al. Linkage of the gene for an X-linked mental retardation disorder to a hypervariable $(AGAT)_n$ repeat motif within the human HPRT locus (Xq26) (submitted)

Koschinsky ML, Beisiegel U, Henne-Bruns D, Eaton DL, Lawn RM (1990) Apolipoprotein(a) size heterogeneity is related to variable number of repeat sequences in its mRNA. Biochemistry 29:640–644

Levinson G, Gutman GA (1987) High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acids Res 15:5323–5338

Lindahl G, Gersdorf E, Menzel HJ, Seed M, Humphries S, Utermann G (1990) Variation in the size of human apolipoprotein(a) is due to a hypervariable region in the gene. Hum Genet 84:563–567

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:397–401

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, et al (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616–1622

Nanda I, Deubelbeiss C, Guttenbach M, Epplen JT, Schmid M (1990) Hetrogenetities in the distribution of $(GACA)_n$ simple repeats in the karyotypes of primates and mouse. Hum Genet 85:187–194

Orita M, Suzuki Y, Sekiya T, Hayashi K (1990) Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. Genomics 5: 874–879

Patel PI, Nussbaum RL, Framson PE, Ledbetter DH, Caskey CT, Chinault AC (1984) Organization of the HPRT gene and related sequences in the human genome. Somatic Cell Mol Genet 10:483–493

Riley J, Butler R, Ogilvie D, Finniear R, Jenner D, Powell S, Anand R, et al (1990) A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. Nucleic Acids Res 18:2887–2890

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, et al (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487–491

Sambrook J, Fritsch EF, Maniatis T (eds) (1989) Molecular cloning: a laboratory manual, 2d ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Savatier P, Trabuchet G, Faure C, Chebloune Y, Gouy M, Verdier G, Nigon VM (1985) Evolution of the primate β-globin gene region: high rate of variation in CpG dinucleotides and in short repeated sequences between man and chimpanzee. J Mol Biol 182:21–29

Sensabaugh GF (1982) Biochemical markers of individuality. In: Saferstein R (ed) Forensic science handbook. Prentice-Hall, Englewood Cliffs, NJ, pp 338–415

Sinnett D, Deragon J-M, Simard LR, Labuda D (1990) Alumorphs: human DNA polymorphisms detected by polymerase chain reaction using Alu specific primers. Genomics 7:331–334

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, et al (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679

Swallow DM, Gendler S, Griffiths B, Corney G, Taylor-Papadimitriou J, Bramwell ME (1987) The human tumor-associated epithelial mucins are coded by an expressed hypervariable gene locus PUM. Nature 328:82–84

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17:6463–6471

Weber JL (1990) Informativeness of human $(dC\text{-}dA)_n$-$(dG\text{-}dT)_n$ polymorphisms. Genomics 7:524–530

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396

Wharton K, Yedvobnick B, Finnerty V, Artavanis-Tsakonas S (1985) opa: a novel family of transcribed repeats shared by the Notch locus and other developmentally regulated loci in *D. melanogaster*. Cell 40:55–62

Wijmenga C, Frants RR, Brouwer OF, Moerer P, Weber JL, Padberg GW (1990) Location of facioscapulohumeral muscular dystrophy gene on chromosome 4. Lancet 336: 651–653

Zuliani G, Hobbs HH (1990) A high frequency of length polymorphisms in repeated sequences adjacent to Alu sequences. Am J Hum Genet 46:963–969